Catalogues as Data: Interpretable NLP Pipelines for Ottoman-Turkish Bibliographies

Mark J. Hill, Ayse Bulus, and Paul Spence
King's College London
Strand, WC2R 2LS
London, UK
Correspondence: mark.j.hill@kcl.ac.uk

Abstract

Bibliographies are both humanities infrastructure and historic record. To computationally analyse them, however, requires implementing complex digitisation and standardisation decisions. This paper turns to Seyfettin Özege's *Eski Harflerle Basılmış Türkçe Eserler Kataloğu* as an example, a scanned set of volumes marked by complex page layouts, degraded typography, irregular entry structures, and historically contingent inconsistencies. With this we present a pipeline that constructs a structured, machine-readable, and analysable dataset out of the 27,000 entries with computer vision, OCR, large and visual language models, sequence-based validation, and custom review tools. This process captures 97.8% of records, with remaining cases capable of being addressed by targeted review. This process demonstrates that combining LLMs with interpretable, review-centric pipelines, offers an appropriate approach for historically complex bibliographic sources.

1 Introduction

Digitisation has increased access to historical bibliographic resources. However, while scanned catalogues are accessible, they do not easily support large-scale querying, aggregation, or complex analysis. This matters as, for many Digital Humanities (DH) researchers, a goal is to interpret historically situated bibliographic structures at scale.

This paper addresses these challenges through Mehmet Seyfettin Özege's *Eski Harflerle Basılmış Türkçe Eserler Kataloğu*, a five-volume bibliography of Ottoman-Turkish printed works (Özege, 1971–1979). Compiled by a single scholar over several decades, the catalogue is both a valuable reference and historical artefact, shaped by uneven access to sources and pragmatic bibliographic choices. These contingencies manifest in non-uniform entry structures which are interesting in their own right, but pose difficulties for computational analysis.

In the following we present a computational pipeline designed to both clean or normalise the catalogue (making its structure computationally legible) while preserving its historical contingencies. To do this we combine computer vision, optical character recognition (OCR), and both Large and Vision Language Models (LLMs and VLMs). Throughout we foreground interpretability, provenance, and scholarly caution rather than automation through a process of extraction (layout handling, entry segmentation, OCR, initial parsing), validation (sequence checks, duplicate detection), and review (implemented through specialised interfaces and tracked operation logs). This framing treats uncertainty and variation as an output: not errors to eliminate, but signals that must be assessed by an expert and are analysable in their own right.

2 Related Work

This work sits within a "catalogues-as-data" orientation in DH, which treats bibliographic catalogues as large-scale datasets, themselves reflections of historical material and practice (Gooding et al., 2025). Examples of this approach can be seen in the work done by the Helsinki Computational History Research Group, who have developed pipelines for harmonising and enriching bibliographic resources, and linking them to full-text corpora. This work has demonstrated how Machine-Readable Cataloguing (MARC) can be transformed into relational datasets suitable for quantitative historical analysis (Hill et al., 2019; Lahti et al., 2019; Tolonen et al., 2021). They have also shown how careful cleaning and unification supports novel downstream historical work, ranging from social network analysis to studying the economics of the book trade (Hill et al., 2023; Tolonen et al., 2025; Tiihonen et al., 2024).

Proceedings of SIGHUM 2026, pages 128–134
March 28-29, 2026 ©2026 Association for Computational Linguistics

This paper differs, however. Rather than starting with MARC records it begins with a scanned bibliography, requiring a multistage pipeline to convert structurally complex facsimiles of pages into individual entries. That is, parsing the interaction of page layout, segmentation, OCR noise, and irregular entry structure without any pre-existing structural encoding comparable to MARC 21. To do this we turn to LLMs, arguing that in these contexts they enable structural interpretation not achievable through rule-based parsing alone.

We are not the first to use LLMs in DH work, with others arguing they can be integrated into mixed-methods workflows to scale qualitative analysis while preserving replicability and transparency (Karjus, 2025). With regard to bibliographic work, recent research has looked at using LLMs to augment existing catalogues (Korpet and Rees, 2025), using them to create structured MARC 21 records (Taniguchi, 2024; Aycock, 2025), and difficulties verifying LLM-generated bibliographic metadata (Kohút et al., 2025; Frenzel, 2025).

This work also sits within Digital Ottoman Studies, a field that in recent years has foregrounded work focused on quantitative and textual analysis of the Ottoman world (Barakat and Yaycıoğlu, 2022). This includes the challenges born out of Ottoman Turkish as a low resource language (Karagöz et al., 2024). An overview of these issues can be found in Yüksek (2022), however central is the fact that Ottoman Turkish texts are inconsistently converted from Arabic script into the Latin alphabet. The same word can be spelled in different ways, which humans can recognise as equivalent, but computers mistake. Due to these language constraints, researchers have advocated for computational workflows in which the researcher is central, rather than fully automated solutions (Aladağ, 2020, 2021, 2024; Kırmızıaltın et al., 2022). Our pipeline extends this orientation to bibliographic data, applying similar principles of validation, traceability, and interpretive control.

3 Background: The Özege Catalogue

Özege's catalogue documents Ottoman-Turkish books and pamphlets excluding newspapers and periodicals. It was conceived as a retrospective scholarly reconstruction rather than a contemporary publishing record, and is based (wherever possible) on direct examination of physical copies.

Özege rejected rigid adherence to bibliographic standards when these conflicted with clarity or usability. As a result, entries vary considerably in completeness, ordering, and expression. For example, dates appear in multiple calendar systems; place and personal names vary in spelling; reprint or edition details are often implicit; and numerous typographic errors exist, including misnumbered and missing entry numbers. These features complicate computational processing.

Each entry is a bounded textual block surrounded by white space. The block corresponds to a single bibliographic item or a work with multiple fascicles or volumes. In very broad terms, each entry loosely follows this ordering: title line(s) in bolded all-caps; an authorship or responsibility statement bolded; publication statement(s) including place of publication and publisher; physical description including number of pages and dimensions; additional notes; and a catalogue number. However, not all components are present, and order is not strictly enforced. For example, fields may be omitted, repeated, or distributed across multiple lines. Some blocks encode multiple editions (reprints, different printers, fascicles) under one heading. Moreover, individual elements may be role-ambiguous, with the same name or phrase functioning as authorial attribution, publisher information, or title continuation depending on context (for examples see Appendix A). Certain elements, such as cross-references and fascicle groupings, also function relationally, linking entries or manifestations rather than describing a single, self-contained record. Therefore, the entry boundary is defined primarily by typographic cues rather than semantic completeness (in particular, white space separation, capitalisation, bold typefaces, and a terminal catalogue number). A level of interpretation is necessary for parsing records, a task for which LLMs are well suited.

While these variations may impact usability in quantitative contexts, they can also be indicative of historical meaning, evidencing changing conventions, uncertain attribution, or uneven bibliographic access. Any NLP-based intervention must, therefore, negotiate the tension between structure and irregularity, extraction and interpretation. While structured outputs are the goal, methodologically we aim to preserve uncertainty via provenance links to page images, and note historically meaningful variation as an output rather than error to be normalised.

4 Pipeline Overview

This paper makes three contributions. First, we present a pipeline that combines vision models, OCR, instruction-tuned LLMs, and sequence-based validation to recover structured bibliographic records from a scanned catalogue without pre-existing markup. Second, we introduce an iterative, anchor-based vision-language extraction strategy that exploits catalogue sequence constraints and spatial context to repair OCR and segmentation failures, achieving improved coverage with minimal false positives. Third, we demonstrate how problem-specific review tools with full operation logging can aid expert intervention in large historical datasets while preserving provenance, reversibility, and interpretive caution. In practice, our pipeline converts scanned PDF volumes into structured bibliographic records through multiple stages. The design prioritises transparency and explicit handling of uncertainty. Each stage produces intermediate artefacts that can be inspected independently, supporting both debugging and scholarly verification. The pipeline proceeds as follows (for a diagram see Appendix B):

Page and column segmentation using projection-based heuristics and a custom Python script to handle two-page, multi-column layouts. This outputs an initial 2,457 pages.

Entry detection via a YOLOv8 model trained on 233 columns covering 1,328 entries, and producing approximately 27,000 individual entry images (Jocher et al., 2023).

OCR extraction using Tesseract and a Turkish language model (Smith, 2007).

Text structured parsing using a pre-trained, instruction-tuned LLM (Qwen3-4B) prompted to interpret OCR output according to the catalogue's bibliographic conventions (Yang et al., 2025).

Validation in which overall catalogue order and comprehensiveness is checked against both the original OCR and the LLM's exported records. This identifies gaps and misnumbered entries.[1]

Vision-language repair using a pre-trained VLM (Qwen3-VL-8B) prompted to re-extract catalogue identifiers from entry images flagged in the previous step (Bai et al., 2025). This step improves coverage and minimises the number of entries which need to be assessed by humans.[2]

Reconstruction and validation is achieved with an interactive tool to assess entries against metadata, including catalogue-number sequence checks, duplicate detection, identifying issues ranging from errors in the original document (mislabelled entries) to entry segments capturing multiple entries requiring further splitting.

Structured LLM parsing of reconstructed, trustworthy entries was achieved using the finalised and verified data, producing a tabular output suitable for downstream scholarly analysis and export.

5 Discussion and Current Results

The pipeline extracted 27,054 candidate entry images via automated segmentation. Using OCR-based parsing alone, reliable catalogue identifiers could be recovered for approximately 85% of entries (assessed based on expected catalogue order and sequencing). Failures were caused by OCR issues (faint numerals, degraded typography, digit substitution errors), segmentation issues, and errors in the original text. By introducing iterative vision-language extraction anchored to catalogue-number sequences, overall identifier recovery increased to 97.8%. While this indicates that VLMs may be a superior tool to OCR at some tasks, resource issues limit their usability for large datasets.

The final 2.2% of obvious errors were assessed through a hybrid workflow that used automated detection to allow for targeted human review. These were identified through sequence breaks, duplicate identifiers, implausible jumps, or spatial inconsistencies. Importantly, the automated process meant manual intervention was only necessary for hundreds rather than thousands of entries, reducing review scope substantially.

Error analysis shows that failures were overwhelmingly attributable to boundary material (prefatory pages, indices), segmentation issues, and genuinely ambiguous cases where catalogue numbering itself is inconsistent or absent in the source.

Manual interventions were carried out using review tools that preserved spatial context (showing neighbouring entries, page layout, and original scans) and supported domain-specific operations such as splitting under-segmented entries, merging over-segmented ones, resolving duplicate identifiers via ID correction, and marking non-entry material (e.g., headers, indices) as garbage. All operations are recorded as structured JSON logs,

[1] Although Qwen3 is a multilingual LLM, a model trained specifically on Turkish, or even Ottoman-Turkish, language could show improvements (Acikgoz et al., 2024).

[2] Ambiguities due to segmentation (e.g., entries spanning columns/pages) are flagged at this stage to be reparsed.

Metadata Field	Coverage
Title	96.5%
Date	88.3%
Pages	88.4%
Dimensions	90.2%
Place	86.6%
Publisher	87.2%
Author	60.9%
Translator	6.0%
Notes	57.7%

Table 1: Coverage of extracted metadata fields across the catalogue (n = 26,712).

enabling full auditability, undo/redo, and reproducibility. Original data is never deleted; instead, interpretive decisions are layered on top of preserved source material. The resulting dataset is therefore not only machine-readable, but traceable: each structured record can be linked back to its source image, OCR text, model outputs, and human decisions.

The resulting corpus exhibits high overall metadata completeness (see Table 1). Lower coverage for authorship (60.9%) and translators (6.0%) reflects the catalogue's original structure rather than extraction failure, as these fields are frequently absent or implicit in the source material.

To assess extraction accuracy we evaluated the pipeline against a manually transcribed subset of 109 entries, independently keyed by a domain expert from the original catalogue. Catalogue identifier extraction achieves 99.1% recall, correctly recovering 108 of 109 identifiers.

Metadata field accuracy was evaluated for the 108 entries with matching catalogue identifiers (see Table 2).[3] Because the pipeline and the gold standard differ in schema granularity and formatting conventions, we report accuracy under multiple matching criteria: exact string match, normalised match (after Unicode normalisation, Turkish-aware case folding, and whitespace collapsing), and containment match (where the gold-standard value is a substring of the extracted value, or vice versa). For dates, we additionally compare extracted four-digit year tokens, since the pipeline outputs combined Hicri and Miladi dates (e.g., "1321 (1903)") whereas the gold standard records them separately.

These results reveal three distinct and systematic sources of mismatch. First, orthographic normalisation: titles achieve 0% exact precision due to the interaction of multiple systematic factors. The catalogue records titles in all-caps while the gold standard uses title case. Lowercasing alone resolves 17% of entries, and a further 24% are recovered by Unicode normalisation and diacritic handling (e.g., 'MUAHEDENAMESİ' vs. 'Muahedenamesi'). Containment matching recovers an additional 25%, accommodating cases where the pipeline captures a truncated or expanded form of the title. For place names, the dominant discrepancy is İstanbul/Istanbul variation, which normalised matching resolves (precision from 66.7% to 86.7%, with 100% precision under fuzzy matching at a 0.85 threshold).

Second, field boundary differences: the pipeline systematically prepends place of publication to the publisher field as this is how it is recorded in the original material (e.g., "Istanbul Orhaniye Matbaası" vs. "Orhaniye Matbaası"), and occasionally absorbs authorial or subtitle information into adjacent fields. This explains why publisher containment precision reaches 80.4%. In most cases the core publisher name is correctly extracted but packaged with additional context. While both the LLM and VLM were prompted with this structural information, and successfully extracted the place data, they failed to purge it from the publisher field. Similarly, the author field sometimes captures the original author or translator alongside or instead of the primary author, reflecting a genuine ambiguity in the catalogue's entry structure that the gold standard resolves through domain expertise and separate columns (e.g., "Name of the author" vs. "Original author").

Third, formatting conventions: the date field achieves only 10.3% normalised precision because the pipeline combines calendar systems into a single string. Under containment matching (which checks whether the Miladi year appears within the extracted date) precision rises to 92.8%, confirming that the relevant temporal information is reliably captured.

All three of these issues can be improved through further refinement, although will require careful schema planning balancing the original source with any desired harmonisation. They are well-defined tasks rather than fundamental limitations of the extraction approach. This suggests that the pipeline's core extraction capability is stronger than exact-match metrics alone suggest.

Overall, these results demonstrate that combining LLMs and VLMs with explicit validation

[3]The evaluation reported here was conducted on raw pipeline output prior to any manual corrections, and therefore represents a lower bound on achievable accuracy.

Field	Exact Precision / Recall	Normalised Precision / Recall	Containment Precision / Recall	Mean similarity
Title	0.0% / 0.0%	41.5% / 40.7%	66.0% / 64.8%	0.82
Author	42.2% / 41.5%	43.8% / 43.1%	60.9% / 60.0%	0.81
Place	66.7% / 49.0%	86.7% / 63.7%	86.7% / 63.7%	0.98
Date	9.3% / 9.3%	10.3% / 10.3%	92.8% / 92.8%	0.64
Publisher	34.0% / 33.7%	36.1% / 35.7%	80.4% / 79.6%	0.86

Table 2: Metadata extraction accuracy on manually transcribed subset (n = 108), under exact, normalised, and containment matching. Mean similarity is the average character-level similarity for entries.

constraints and review-centric tooling enables reliable extraction from historically irregular bibliographic sources. Identifier recovery is near-complete (99.1%), and metadata fields show high containment accuracy (61–93%), indicating that the pipeline captures relevant information even where exact string rendering differs from the gold standard. The dataset in its current form is usable for analyses that tolerate orthographic variation, such as temporal and geographic distributions of Ottoman-Turkish publishing. Precision-sensitive tasks, including author network analysis, will benefit from the planned normalisation and disambiguation work outlined in Section 6.

6 Conclusion

This case study illustrates broader methodological lessons for NLP in DH contexts. First, task decomposition matters: computer vision, OCR, and LLMs each excel at different stages. Our experience suggests that no single model equally handles all tasks, and task decomposition improves accuracy, interpretability, and refinement. Second, explanation and traceability are as important as extraction accuracy when outputs are intended for scholarly use.

More broadly, the pipeline reframes LLMs not as replacements for bibliographic expertise, but as tools for scaling interpretive labour while maintaining epistemic grounding. By explicitly encoding uncertainty and preserving historical variation, we align computational processing with the values of humanistic inquiry rather than subordinating them to optimisation alone.

Future work includes further harmonisation as well as disambiguation. In the former case, this includes names, places, and dates to start. In the latter, many entries catalogue multiple individual editions which could be disambiguated into individual records, expanding the bibliography substantially. With this data, researchers will be able to further explore the bibliography and history of Ottoman print culture from temporal, geographic, and more generally quantitative perspectives.

Limitations

This work has limitations. First, while catalogue identifier recovery is near-complete, metadata extraction remains sensitive to OCR noise and orthographic variation. Second, although manual intervention is limited in scale, it is still required, and requires some level of expertise and time. Third, the pipeline has been evaluated on a single, albeit complex, bibliographic source; generalisation to other catalogues with different typographic conventions or languages would require retraining segmentation models and adapting prompts and validation rules. The abstract pipeline, however, can be taken as a valid starting point and in some instances could be re-used successfully (e.g., sequence validation logic, reviewing architecture). Fourth, while Qwen3 is a multilingual LLM (including the Turkic language family), more work needs to be done to assess potential errors in terms of low-resource languages. Finally, computational constraints limited the use of VLMs to selected cases rather than full-corpus application, which may cap achievable recall in the most degraded instances.

Acknowledgments

The authors would like to acknowledge King's College London and the Computational Research, Engineering and Technology Environment (CREATE) for providing access to high performance computing resources. Additionally, we would like to thank the peer reviewers whose close reading allowed for both clarifications and highlighted an error in a previously reported metric.

References

Emre Can Acikgoz, Mete Erdogan, and Deniz Yuret. 2024. Bridging the bosphorus: Advancing turkish large language models through strategies for

low-resource language adaptation and benchmarking. *Preprint*, arXiv:2405.04685.

Fatma Aladağ. 2020. Deciphering Ottoman Turkish manuscripts with LexiQamus. The Digital Orientalist.

Fatma Aladağ. 2021. Innovative designs on Ottoman Turkish search engines: Wikilalala and Müteferriqa. The Digital Orientalist.

Fatma Aladağ. 2024. Exploring the (digital) world of Ottoman Turkish texts: The digital Ottoman corpora. The Digital Orientalist.

Mary Aycock. 2025. Prompting generative ai to catalog: The promise and the reality. *College Research Libraries News*, 86(10):423.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

N. E. Barakat and A. Yaycıoğlu. 2022. Critical approaches to digital ottoman studies: Promises and challenges. *Journal of the Ottoman and Turkish Studies Association*, 9(2):17–31.

Fran Frenzel. 2025. Smart enough to mislead: The functional shortcomings and ethical dilemmas of generative ai use in metadata work. *Catalogue & Index*, (211). Received 10 June 2025; Published 17 June 2025.

Paul Gooding, Melissa Terras, and Sarah Ames, editors. 2025. *Library Catalogues as Data: Research, Practice and Usage*. Facet Publishing.

Mark J. Hill, Ville Vaara, Tanja Säily, Leo Lahti, and Mikko Tolonen. 2019. Reconstructing intellectual networks: From the ESTC's bibliographic metadata to historical material. In *CEUR Workshop Proceedings*.

Mark J. Hill, Ville Vaara, and Mikko Tolonen. 2023. Communication and idea transmission across historical communities: A quantitative analysis of early modern nonconformist networks. *Huntington Library Quarterly*, 86(2):377–407.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Ultralytics yolov8. Version 8.x.

Fatih Karagöz, Berat Doğan, and Şaziye Betül Özateş. 2024. Towards a clean text corpus for Ottoman Turkish. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 62–70, Bangkok, Thailand and Online. Association for Computational Linguistics.

Andres Karjus. 2025. Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence. *Humanities and Social Sciences Communications*, 12:277.

Süphan Kırmızıaltın, Fatma Aladağ, and Elif Derin Can. 2022. Crowdsourcing Ottoman cultural heritage: OTurC and participatory digital corpora creation. *Journal of the Ottoman and Turkish Studies Association*, 9(2):37–42.

Jan Kohút, Martin Dočekal, Michal Hradiš, and Marek Vaško. 2025. Bibliopage: A dataset of scanned title pages for bibliographic metadata extraction. *Preprint*, arXiv:2503.19658.

Sheldon Korpet and Nathalie Rees. 2025. Augmenting cataloguers: planning an AI agent to generate MARC21 records. *Catalogue & Index*, (211). Received 2 June 2025; published 17 June 2025.

Leo Lahti, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. 2019. Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly*.

Ray Smith. 2007. An overview of the tesseract ocr engine. *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633.

Shoichi Taniguchi. 2024. Creating and evaluating MARC 21 bibliographic records using ChatGPT. *Cataloging & Classification Quarterly*, 62(5):527–546.

Iiro Tiihonen, Leo Lahti, and Mikko Tolonen. 2024. Print culture and economic constraints: A quantitative analysis of book prices in eighteenth-century britain. *Explorations in Economic History*, 94:101614.

Mikko Tolonen, Mark J. Hill, Ahmed Z. Ijaz, Ville Vaara, and Leo Lahti. 2021. Examining the early modern canon: The english short title catalogue and large-scale patterns of cultural production. In Ian Baird, editor, *Data Visualization in Enlightenment Literature and Culture*. Palgrave Macmillan, Cham.

Mikko Tolonen, Mark J. Hill, and Jani Marjanen. 2025. Library catalogues as data. In Lara Goodings and Melissa Terras, editors, *Book Formats, Printing Practices and Reading Habits in Early Modern Europe*. UCL Press, London. Book chapter.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ahmet Yusuf Yüksek. 2022. Natural language processing in ottoman turkish: Opportunities and drawbacks. PhD Candidate, New York University.

Mehmet Seyfettin Özege. 1971–1979. *Eski Harflerle Basılmış Türkçe Eserler Kataloğu*. Fatih Yayınevi, İstanbul. 5 vols.

A Appendix

Two-page Spread

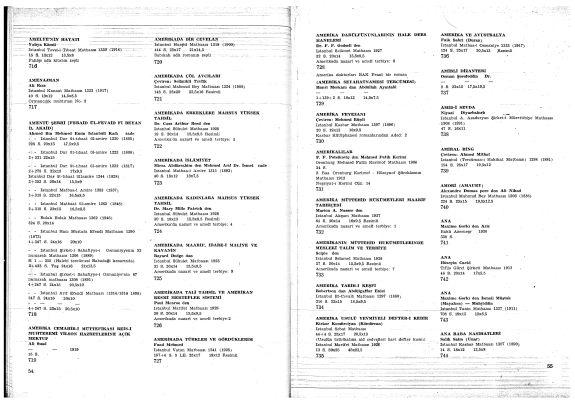

Figure 1: Example two-page spread including catalogue entry 717.

Computationally Segmented Entry

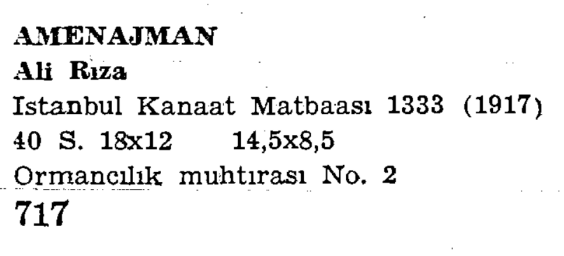

Figure 2: Example of computationally segmented entry (717).

Raw OCR

```
AMENAJMAN İ
Ali Rıza
Istanbul Kanaat Matbaası 13833 (1917)
40 S. 18x12 14,5x8,5
, Ormancılık muhtırası No. 2
1717
```

Parsed Metadata

Title: AMENAJMAN İ
Author: Ali Rıza
Translator: NA
Publisher: Istanbul Kanaat Matbaası
Place: İstanbul
Date: 13833 (1917)
Pages: 40 S.
Dimensions: 18x12, 14,5x8,5
Notes: Ormancılık muhtırası No. 2
Catalogue number: 717

Note, this example highlights how VLM re-assessment of computationally identified catalogue errors is able to correct OCR errors (in this case, catalogue number 1717 is corrected to 717). However, other OCR errors persist and are re-assessed in the final LLM parsing stage.

B Appendix

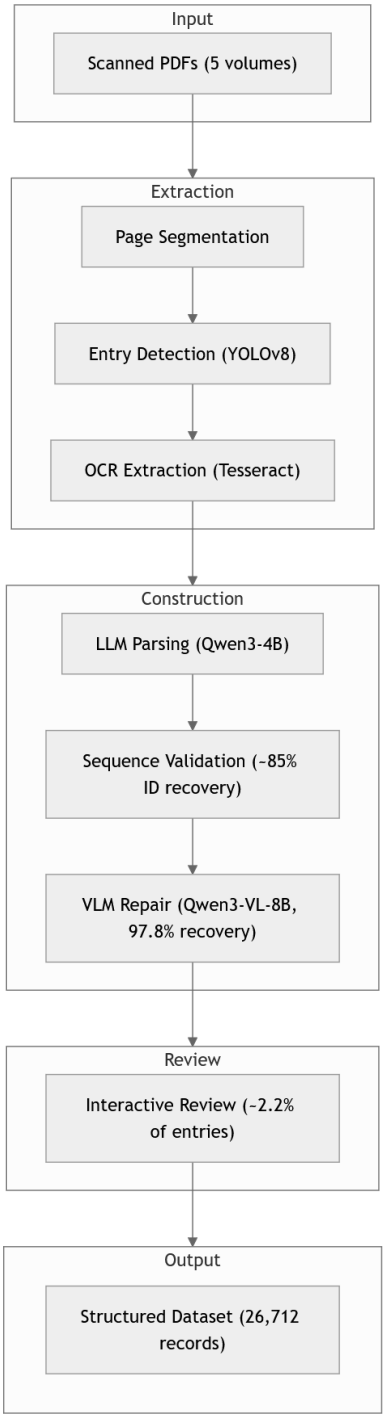

Figure 3: Pipeline overview